# 大数据管理方法与应用第三次作业

大数据 001

郅啸淇

学号：2184114639

2023 年 4 月 16 日

# 1 抛硬币的后验分布

## 1.1 分布与图像

由课堂证明已知，抛硬币的后验分布为 $P(\theta|x) = Beta(\theta|a+x, b+n-x)$，将不同参数的各组数据分布情况绘图如图 1 所示。

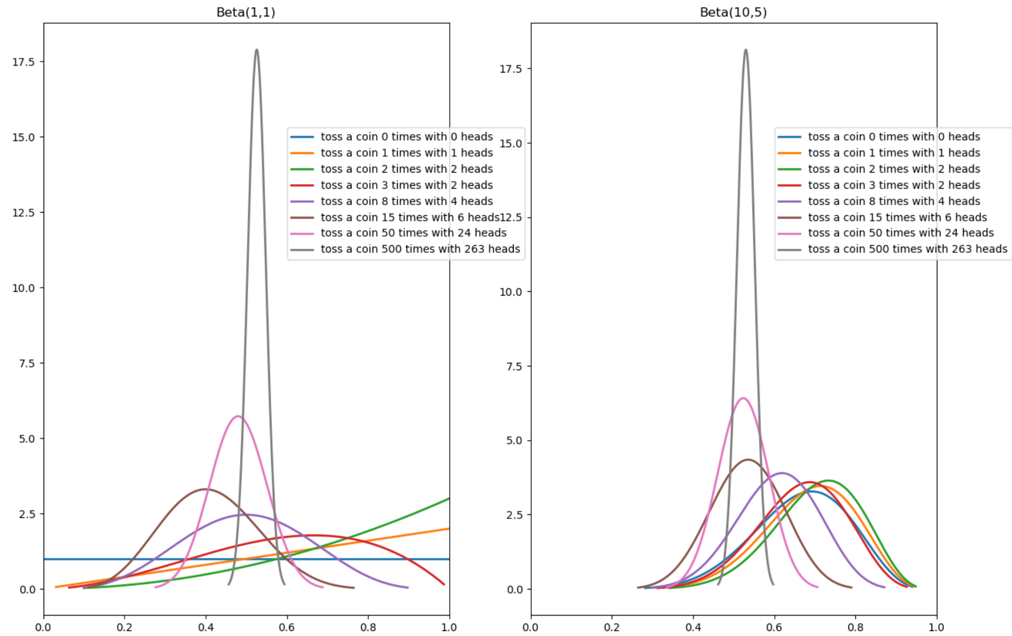


图 1: 分布情况

## 1.2 完整代码

```
import numpy as np
from scipy.stats import beta
import matplotlib.pyplot as plt

plt.rcParams['axes.unicode_minus'] = False
n_list = [0, 1, 2, 3, 8, 15, 50, 500]
x_list = [0, 1, 2, 2, 4, 6, 24, 263]


def draw(ax, a, b, n, x):
    post_a = a + x
    post_b = b + n - x
    x_line = np.linspace(beta.ppf(0.001, post_a, post_b), beta.ppf(0.999, post_a,
        post_b), 1000)
    ax.plot(x_line, beta.pdf(x_line, post_a, post_b), lw=2,
            label="toss a coin {n} times with {x} heads". format(n=n, x=x))
    ax.legend(loc=(0.6,0.6))
    plt.xlim(0, 1)


plt.figure(figsize=(15, 10))
ax = plt.subplot(1, 2, 1)
ax.set_title("Beta(1,1)")
for n, x in  zip(n_list, x_list):
    draw(ax, 1, 1, n, x)
ax = plt.subplot(1, 2, 2)
ax.set_title("Beta(10,5)")
for n, x in  zip(n_list, x_list):
    draw(ax, 10, 5, n, x)
plt.show()
```

# 2 共轭先验的证明

## 2.1 证明多项分布的共轭先验是狄利克雷分布

似然函数为多项分布，其中 $\theta_i$ 代表第 i 类出现的概率，$n_i$ 代表第 i 类出现的次数。

多项分布的先验分布 $P(x|\theta) = \frac{n!}{n_1!n_2!n_3!n_4!...n_k!}\prod_{i=1}^{k}\theta_i^{n^i}$,其中 $\sum_{i=1}^{k}\theta_i = 1$

使用 Gamma 函数对阶乘进行近似，有 $\Gamma(x+1) = x!$，则 $P(x|\theta) = \frac{\Gamma(n+1)}{\sum_{i=1}^{k}\Gamma(n_i+1)}\prod_{i=1}^{k}\theta_i^{n_i}$

假设概率 $\theta = (\theta_1, \theta_2, ..., \theta_k)$ 的先验分布为参数是 $\alpha = (\alpha_1, \alpha_2, \alpha_3, ..., \alpha_k)$ 的狄利克雷分布 $Dir(\alpha)$

有 $P(\theta) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k}\theta_i^{\alpha_i-1}$

计算

$$
\begin{aligned}
P(x) &= \int P(x|\theta)P(\theta)d\theta \\
&= \int_0^1 \frac{\Gamma(n+1)}{\prod_{i=1}^{k}\Gamma(n_i+1)}\prod_{i=1}^{k}\theta_i^{n_i}\frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k}\theta_i^{\alpha_i-1}d\theta \\
&= \frac{\Gamma(n+1)\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(n_i+1)\Gamma(\alpha_i)}\frac{\prod_{i=1}^{k}\Gamma(n_i+\alpha_i)}{\Gamma(\sum_{i=1}^{k}n_i+\alpha_i)}\int_0^1\frac{\Gamma(\sum_{i=1}^{k}n_i+\alpha_i)}{\prod_{i=1}^{k}\Gamma(n_i+\alpha_i)}\prod_{i=1}^{k}\theta_i^{\alpha_i+n_i-1}d\theta \\
&= \frac{\Gamma(n+1)\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(n_i+1)\Gamma(\alpha_i)}\frac{\prod_{i=1}^{k}\Gamma(n_i+\alpha_i)}{\Gamma(\sum_{i=1}^{k}n_i+\alpha_i)}\int_0^1 Dir(n+\alpha)d\theta \\
&= \frac{\Gamma(n+1)\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(n_i+1)\Gamma(\alpha_i)}\frac{\prod_{i=1}^{k}\Gamma(n_i+\alpha_i)}{\Gamma(\sum_{i=1}^{k}n_i+\alpha_i)}
\end{aligned}
$$

然后计算 $\theta$ 的后验分布

$$
\begin{aligned}
P(\theta \mid x) &= \frac{P(x \mid \theta)P(\theta)}{P(x)} \\
&= \frac{\frac{\Gamma(n+1)}{\prod_{i=1}^{k}\Gamma(n_i+1)}\prod_{i=1}^{k}\theta_i^{n_i}\frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k}\theta_i^{\alpha_i-1}}{\frac{\Gamma(n+1)\Gamma(\sum_{i=1}^{k}\alpha_i)}{\Gamma(\sum_{i=1}^{k}n_i+\alpha_i)}\prod_{i=1}^{k}\frac{\Gamma(n_i+\alpha_i)}{\Gamma(n_i+1)\Gamma(\alpha_i)}} \\
&= \frac{\Gamma(\sum_{i=1}^{k}n_i+\alpha_i)}{\prod_{i=1}^{k}\Gamma(n_i+\alpha_i)}\prod_{i=1}^{k}\theta_i^{n_i+\alpha_i-1} \\
&= Dir(n+\alpha)
\end{aligned}
$$

得到 $\theta$ 先验分布和后验分布均为狄利克雷分布，且似然函数为多项分布，故多项分布的共轭先验为狄利克雷分布

## 2.2 证明泊松分布的共轭先验为伽马分布

似然函数是指数分布，n 代表事件发生次数，$\lambda$ 代表单位时间内随机事件的平均发生次数

则有先验分布 $P(x=n \mid \lambda) = \lambda e^{-\lambda x}$

假设 $\lambda$ 服从参数为 $(a,b)$ 的伽马分布 $P(\lambda) = \frac{\lambda^{a-1}e^{-b\lambda}b^a}{\Gamma(a)}$

计算 $P(x)$

$$
\begin{aligned}
P(x) &= \int P(x \mid \lambda)P(\lambda)d\lambda \\
&= \int_0^1 \lambda e^{-\lambda x}\frac{\lambda^{a-1}e^{-b\lambda}b^a}{\Gamma(a)}d\lambda \\
&= \frac{b^a}{\Gamma(a)}\frac{\Gamma(a+1)}{(b+1)^{a+1}}\int_0^1 \lambda^a e^{-(b+1)\lambda}\frac{(b+1)^{a+1}}{\Gamma(a+1)}d\lambda \\
&= \frac{b^a}{\Gamma(a)}\frac{\Gamma(a+1)}{(b+1)^{a+1}}\int_0^1 Ga(a+1,b+1)d\lambda \\
&= \frac{b^a}{\Gamma(a)}\frac{\Gamma(a+1)}{(b+1)^{a+1}}
\end{aligned}
$$

得到 $\lambda$ 的后验分布

$$\begin{aligned} P(\lambda \mid x) &= \frac{P(x \mid \lambda)P(\lambda)}{P(x)} \\ &= \frac{\lambda e^{-\lambda x}\frac{\lambda^{a-1}e^{-b\lambda}b^a}{\Gamma(a)}}{\frac{b^a}{\Gamma(a)}\frac{\Gamma(a+1)}{(b+1)^{a+1}}} \\ &= \frac{(b+1)^{a+1}}{\Gamma(a+1)}\lambda^a e^{-(b+1)\lambda} \\ &= Ga(a+1, b+1) \end{aligned}$$

$\lambda$ 的先验和后验分布都是伽马分布，似然函数是泊松分布，所以泊松分布的共轭先验是伽马分布

## 2.3 证明方差已知的正态分布的共轭先验是正态分布

似然函数是方差已知的正态分布，有 $P(x \mid \mu) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
假设 $\mu$ 服从参数为 $(a, b^2)$ 的正态分布 $\mu \sim N(a, b^2)$，则有
$P(\mu) = \frac{1}{b\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\mu-a}{b}\right)^2}$

计算 $P(x)$

$$
\begin{aligned}
P(x) &= \int P(x \mid \mu) P(\mu) d\mu \\
&= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \frac{1}{b\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\mu-a}{b}\right)^2} d\mu \\
&= \frac{1}{2\sigma b\pi} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\frac{(x-\mu)^2 b^2 + (\mu-a)^2\sigma^2}{\sigma^2 b^2}} d\mu \\
&= \frac{1}{2\sigma b\pi} \int_{-\infty}^{\infty} exp\left(-\frac{1}{2}\left(\frac{\left(\mu - \frac{xb^2+a\sigma^2}{\sigma^2+b^2}\right)^2}{\frac{\sigma^2 b^2}{\sigma^2+b^2}} + \frac{(x-a)^2}{\sigma^2+b^2}\right)\right) d\mu \\
&= \frac{e^{-\frac{1}{2}\frac{(x-a)^2}{\sigma^2+b^2}}}{2\sigma b\pi} \frac{\sigma b}{\sqrt{\sigma^2+b^2}} \sqrt{2\pi} \int_{-\infty}^{\infty} \frac{1}{\frac{\sigma b}{\sqrt{\sigma^2+b^2}}\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{\left(\mu - \frac{xb^2+a\sigma^2}{\sigma^2+b^2}\right)^2}{\frac{\sigma^2 b^2}{\sigma^2+b^2}}\right)\right) d\mu \\
&= \frac{1}{\sqrt{\sigma^2+b^2}\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-a)^2}{\sigma^2+b^2}} \int_{-\infty}^{\infty} \mu \sim N\left(\frac{xb^2+a\sigma^2}{\sigma^2+b^2}, \frac{\sigma b}{\sqrt{\sigma^2+b^2}}\right) d\mu \\
&= \frac{1}{\sqrt{\sigma^2+b^2}\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-a)^2}{\sigma^2+b^2}}
\end{aligned}
$$

计算 $\mu$ 的后验分布

$$
\begin{aligned}
P(\mu \mid x) &= \frac{P(x \mid \mu) P(\mu)}{P(x)} \\
&= \frac{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \frac{1}{b\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\mu-a}{b}\right)^2}}{\frac{1}{\sqrt{\sigma^2+b^2}\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-a)^2}{\sigma^2+b^2}}} \\
&= \frac{\sqrt{\sigma^2+b^2}}{\sigma b\sqrt{2\pi}} e^{-\frac{1}{2}\left(\left(\frac{x-\mu}{\sigma}\right)^2 + \left(\frac{\mu-a}{b}\right)^2 - \frac{(x-a)^2}{\sigma^2+b^2}\right)} \\
&= \frac{1}{\frac{\sigma b}{\sqrt{\sigma^2+b^2}}\sqrt{2\pi}} e^{-\frac{1}{2}\frac{\left(\mu - \frac{xb^2+a\sigma^2}{\sigma^2+b^2}\right)^2}{\frac{\sigma^2 b^2}{\sigma^2+b^2}}} \\
&= N\left(\frac{xb^2+a\sigma^2}{\sigma^2+b^2}, \frac{\sigma b}{\sqrt{\sigma^2+b^2}}\right)
\end{aligned}
$$

$\lambda$ 的先验和后验分布都是正态分布，似然函数是正态分布分布，所以方差已知的正态分布的共轭先验是正态分布

## 2.4 证明均值已知的正态分布的共轭先验是逆伽马分布

似然函数是均值已知的正态分布，则有 $P(x \mid \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

假设参数 $\sigma^2$ 服从参数为 $(a,b)$ 的逆伽马分布 $\sigma^2 \sim IGa(a,b)$

$p(\sigma^2) = \frac{b^a}{\Gamma(a)} \left(\frac{1}{\sigma^2}\right)^{a+1} e^{-\frac{b}{\sigma^2}}$

计算 $P(x)$

$$
\begin{aligned}
P(x) &= \int P(x \mid \sigma^2) P(\sigma^2) d\sigma^2 \\
&= \int_0^\infty \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \frac{b^a}{\Gamma(a)} \left(\frac{1}{\sigma^2}\right)^{a+1} e^{-\frac{b}{\sigma^2}} d\sigma^2 \\
&= \frac{1}{\sqrt{2\pi}} \frac{b^a}{\Gamma(a)} \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{a+\frac{3}{2}} e^{-\frac{b}{\sigma^2} - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} d\sigma^2 \\
&= \frac{1}{\sqrt{2\pi}} \frac{b^a}{\Gamma(a)} \frac{\Gamma(a+\frac{1}{2})}{\left(b+\frac{1}{2}(x-\mu)^2\right)^{a+\frac{1}{2}}} \int_0^\infty \frac{\left(b+\frac{1}{2}(x-\mu)^2\right)^{a+\frac{1}{2}}}{\Gamma(a+\frac{1}{2})} \left(\frac{1}{\sigma^2}\right)^{a+1+\frac{1}{2}} e^{-\frac{b+\frac{1}{2}(x-\mu)^2}{\sigma^2}} d\sigma^2 \\
&= \frac{1}{\sqrt{2\pi}} \frac{b^a}{\Gamma(a)} \frac{\Gamma(a+\frac{1}{2})}{\left(b+\frac{1}{2}(x-\mu)^2\right)^{a+\frac{1}{2}}} \int_0^\infty \sigma^2 \sim IGa\left(a+\frac{1}{2}, b+\frac{1}{2}(x-\mu)^2\right) d\sigma^2 \\
&= \frac{1}{\sqrt{2\pi}} \frac{\Gamma(a+\frac{1}{2})}{\Gamma(a)} \frac{b^a}{\left(b+\frac{1}{2}(x-\mu)^2\right)^{a+\frac{1}{2}}}
\end{aligned}
$$

计算后验分布

$$
\begin{aligned}
P(\sigma^2 \mid x) &= \frac{P(x \mid \sigma^2) P(\sigma^2)}{P(x)} \\
&= \frac{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \frac{b^a}{\Gamma(a)} \left(\frac{1}{\sigma^2}\right)^{a+1} e^{-\frac{b}{\sigma^2}}}{\frac{1}{\sqrt{2\pi}} \frac{\Gamma(a+\frac{1}{2})}{\Gamma(a)} \frac{b^a}{\left(b+\frac{1}{2}(x-\mu)^2\right)^{a+\frac{1}{2}}}} \\
&= \frac{\left(b+\frac{1}{2}(x-\mu)^2\right)^{a+\frac{1}{2}}}{\Gamma(a+\frac{1}{2})} \left(\frac{1}{\sigma^2}\right)^{a+1+\frac{1}{2}} e^{-\frac{b+\frac{1}{2}(x-\mu)^2}{\sigma^2}} \\
&= IGa\left(a+\frac{1}{2}, b+\frac{1}{2}(x-\mu)^2\right)
\end{aligned}
$$

$\mu$ 的先验和后验分布都是逆伽马函数，似然函数是正态分布分布，所以均值已知的正态分布的共轭先验是逆伽马分布